Specification and evaluation of an assessment engine for educational games

Chaudy, Yaelle; Connolly, Thomas

Published in:
Entertainment Computing

DOI:
10.1016/j.entcom.2018.07.003

Published: 26/07/2018

Document Version
Peer reviewed version

Specification and Evaluation of an Assessment Engine for Educational Games: Empowering Educators with an Assessment Editor and a Learning Analytics Dashboard

Yaëlle Chaudy*, Thomas Connolly
University of the West of Scotland, Paisley UK
yaelle.chaudy@uws.ac.uk, thomas.connolly@uws.ac.uk

Abstract

Assessment is a crucial aspect of any teaching and learning process. Educational games offer promising advantages for assessment; personalised feedback to students and automated assessment process. However, while many teachers agree that educational games increase motivation, learning and retention, few of them are ready to fully trust them as an assessment tool. We believe there are two main reasons for this lack of trust: educators are not given sufficient information about the gameplays, and many educational games are distributed as black-boxes, unmodifiable by teachers. This paper presents an assessment engine designed to separate a game and its assessment. It allows teachers to modify a game's assessment after distribution and visualise gameplay data via a learning analytics dashboard. The engine was evaluated quantitatively by 31 educators. Findings were overall very positive: both the assessment editor and the learning analytics dashboard were rated useful and easy to use. The evaluation also indicates that, having access to EngAGe, educators would be more likely to trust a game's assessment. This paper concludes that EngAGe can be used by educators effectively to modify educational games' assessment and visualise gameplay data, and that it contributes to increasing their trust in educational games as an assessment tool.

Keywords

Educational games; assessment; learning analytics; assessment editor; assessment engine

1. Introduction

Games-based learning (GBL) is increasingly used as a supplementary tool for education. GBL offers a variety of advantages to assist traditional teaching. They can, for instance, allow students to learn at their own pace and they are a safe and controlled environment for students to learn through trial and error. Various institutions use GBL for learning and training, ranging from schools (Kiili & Ketamo, 2017) and higher education institutions (Cózar-Gutiérrez & Sáez-López, 2016) to healthcare (Lv, Esteve, Chirivella, & Gagliardo, 2017; Sliney & Murphy, 2008) and the army (W. L. Johnson, 2007; Zyda, 2005).

However, while many teachers agree that GBL increases motivation towards learning (Sandford, Ulicsak, Facer, & Rudd, 2006) and despite the evidence that games are valid assessment tools (Harteveld & Sutherland, 2015), there seems to be a lack of trust in an educational game's assessment (Sandford et al., 2006; Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2012). Teachers need to feel in control before introducing a new tool in the classroom and there is a need for ownership over the game (Ketelhut & Schifter, 2011); without control, educators might feel threatened by a game rather than supported by it.

One of the main limitations of GBL is that educational games are too often distributed as "*black-boxes*"; they are closed and self-contained systems, making it difficult to modify or retrieve data from (Serrano-Laguna et al., 2017). This can mean that the potential of the game and its attractiveness to educators are

reduced. Indeed, in traditional teaching, improvisation and adaptation to students represent a key aspect of the educator's role (Hunt, 1976), however, teachers tend to lose this capacity with the introduction of a tool they cannot modify to suit the needs of their students. Then, they cannot retrieve data about the gameplays to appreciate whether their teaching goals have been met. Educators and researchers have very little insight about what the students learn through a computer game and how they interact with it. Learning Analytics (LA) is an emerging field based on data mining processes (Siemens & Gasevic, 2012) that can provide such detailed reports about the gameplays; data from the gameplays of several educational games are collected and data mining algorithms allow conclusions to be drawn about the games and the players. However, due to the novelty of the field, presently very few papers exist on LA and its application in GBL and LA is still beyond the reach of most teachers (L. Johnson et al., 2013).

Various platforms such as <e-Adventure> (Torrente, Serrano-Laguna, del Blanco Aguado, Moreno-Ger, & Fernandez-Manjon, 2014) or e-CLIL (Hainey & Connolly, 2013) provide educators with the ability to create and modify their own computer games; <e-Adventure> even includes a learning analytics module(Martinez-Ortiz & Fernandez-Manjon, 2017). These games engines externalise content and assessment integration from the game's code and partially address the problems identified previously. However, these engines were created for educators alone; they are not meant to be used when working with game developers and therefore only provide limited options in terms of game genres and assessment integration. Teachers sometimes lack the time to develop the games themselves or there is a need for a type of game not offered by such platforms.

To summarise, computer games are a powerful tool for learning and assessment but they are often underused by educators, particularly for assessment. We propose three key improvements that could be made for GBL to be more teacher-friendly. First, teachers should be given more control over the game and they should feel a sense of ownership toward the game. Second, the games should be made more flexible, allowing educators to modify and adapt them. The third improvement is the introduction of more detailed reports on the gameplays through LA that will provide teachers with an insight into the appropriateness of the assessment regime used in the game and their students' learning outcomes. It would be optimal to look at all three improvements from a general point of view, addressing assessment integration as well as all other facets of a game such as content integration, story line, graphics and sounds. However, this paper focuses on assessment as it is central in the learning and teaching process.

The aim of this study is to develop and evaluate an assessment engine that would facilitate integrating these three improvements to educational games. In this paper, we present an assessment engine, EngAGe (an Engine for Assessment in Games), that is used by developers during the development of an educational game and it provides tools for educators after distribution of the game. Our approach is based on the externalisation of the assessment. The resulting modularity offers the possibility to modify the assessment logic via an online editor without interfering with the game mechanics and to retrieve information about the gameplays through an LA dashboard.

This paper is divided into five sections as follows. In Section 2, we present a summary of the literature on LA associated with educational games. In Section 3, we explain how EngAGe is used by educators, detailing the design for the assessment editor and the LA dashboard. In Section 4, we present the findings of an evaluation of the tool carried out with 31 educators. Finally, Section 5 draws conclusions and discusses future directions of our research.

2. Previous Research

This section presents the findings of a literature review performed for this research and reviews the different approaches to using LA in GBL. No restriction was imposed on the dates of the papers, however, the oldest relevant study identified was published in 2011 reflecting how recent the topic of LA in games is. The following search terms were used: "*learning analytics"* AND *game*. The search was performed on 15 databases relevant to education, information technology and/or social science: ACM (Association for Computing Machinery), ASSIA (Applied Social Sciences Index and Abstracts), BioMed Central, Cambridge Journals Online, ChildData, Index to Theses, Oxford University Press (journals), Science Direct, EBSCO (consisting of Psychology and Behavioural Science, PsycINFO, SocINDEX, Library, Information Science and Technology Abstracts, CINAHL), ERIC (Education Resources Information Center), IngentaConnect, Infotrac (Expanded Academic ASAP), Emerald, Springer and IEEE (Institute of Electrical and Electronics Engineers) Computer Society Digital Library (CSDL). Relevant papers were identified based on two criteria: papers discussing learning analytics in games and papers presenting a framework for learning analytics in educational games. Papers presenting learning analytics outside of a game environment were excluded. Where possible, the search was based on abstract, titles and keywords to focus on relevant papers. A total of 364 papers were returned published between 2011 and 2016, 22 of these papers were relevant to this review, as summarised in Table 1. These papers are comprised of five book chapters, 14 conference papers and three journal papers. The studies presented in these papers differed in three main aspects: the data collected, the type of analysis applied, and the target users of the tool. These three aspects were categorised and are described in this section. Eleven of the relevant papers were used in real life situations and four showed empirical evidence of the usefulness of the system presented. None presented evidence of its usability.

Table 1: Summary of the literature review on LA in GBL

Study	Description and target users	Data collected	Data Analysis and technology used	Empirical evidence
Information visualisation				
Duval (2011)	Explains how existing tracking and social network services can inspire LA. Goal oriented visualisations. Targets learners and teachers.	Two types of data: The time (total, average per document, time of the day) and numbers (of accessed resources, logins, clicks, artefacts produced, assignments finished)	Visualisation based on Contextualised Attention Metadata and Ontology-based user interaction context models. Line charts, bar charts, parallel coordinates.	n/a
Kickmeier-Rust and Albert (2013)	Presents ProNIFA, a tool for learning analytics in virtual worlds. Aimed at teachers.	Performance data such as test results, activities is collected.	A probabilistic model is used for assessing the student's. The teacher interface displays it with graphs, charts etc.	n/a
Reese (2014)	Presents the CyGaMEs approach to LA and embedded assessment.	The player's variables, actions + achievements are stored every 10sec for the timed report.	Graphs to show the evolution of player's progress. A digital knowledge map for assessment.	Data is used from real gameplays of the CyGaMEs Selene.
Minović and Milovanović (2013) and Minović, Milovanović, Šošević, and González (2015)	Presents a real-time tracking tool of students learning. The tool is mainly aimed at teachers for real-time reaction but can also be used by the players.	Student progress is monitored based on four models: knowledge model, game objects model, Anderson's taxonomy model and learning path model.	Visualisation in the form of a circular graph that represent the whole learning progress of a student. The graph is composed of a centre and three levels of rings.	Experimental study with a group of 6 and 20 students, results indicate that the LA helps educators identify and solve learning problems.
Holman, Aguilar, and Fishman (2013)	Presents the learning analytics of, a gamified learning management system: GradeCraft. Aimed at teachers and students.	Scores, badges earned by player, percentage of completion and final grades are monitored as well as the number of logins + content views.	Progress bars, box-and-whisker plot for score comparison, two-way tables to compare students and line graphs for evolution of score.	Two case studies integrating GradeCraft into a videogame class and a political science course.
Fulantelli, Taibi, and Arrigo (2013)	Describes the LA platform of, a mobile learning environment, MeLOD, for teachers to identify preferred activities and monitor students' progress.	The MeLOD ontology represents all the information stored: User information, configuration of the system, session data, social activity and other activities.	The dashboard uses bar carts, pie charts, tables and visual indicators to display statistical information about the students and the activities.	n/a
(Serrano-Laguna & Fernandez-Manjon, 2014; Serrano-Laguna et al., 2012)	Proposes an LA framework for GBL based on 7 steps: Select, Capture, Aggregate & report, Assess, Use, Use & refine and Share. Aimed at teachers and students.	Personal information about the player, academic information, player's interactions, states and scores.	An aggregation model is used to make sense of the data collected. Teachers author assessment rules. Information displayed with tables, heatmaps, screenshots, graphs etc.	Various case studies have been carried out, testing the tool within <e-adventure>.

Harrer (2013)	Presents the Metafora system, its log-channel (user actions) and analysis-channel (analysis results). Aimed at learners, teachers and researchers.	Four types of data: User actions (in CoLoForm format), Planning maps (created by user), States and Analysis (that can be used for further analysis)	A relational database is used for the actions and planning maps, a nonSQL database for the artefacts produced. An interface offers offline analysis.	Studies undertaken for 7 months with 905 users. There is a replay possibility for evaluation of the feedback generated.
Freire, del Blanco, and Fernandez-Manjon (2014)	Explores the integration of SGs into MOOCs and presents INSIGHTS a general plugin for analytics aimed at students, teachers and researchers	The data collected depends on the system INSIGHT is plugged into	The processing techniques used are not clearly specified; the system seems to rely mostly on visualisation and generation of reports.	n/a
Liu, Lee, Kang, and Liu (2015)	Data visualisation of gameplay data to answer research questions about players' use of a game	Gameplay logs (scores, time spent etc.) and student characteristics (performance, fantasy proneness, game engagement etc.)	Visualisation of data on line charts and tables	Two studies (n= 38 and 64) to determine how different students used the tool.
Herrler, Grubert, Kajzer, Behrens, and Klamma (2016)	Game-editor for teachers to create serious games. The games created collect meaningful data for LA. Aimed at students and teachers.	The game tracked input traces, mouse clicks, events as well as level completion and basic numbers (e.g. total number of gameplays, scores…) and badges.	Data collected using the GLEANER API (Serrano-Laguna, Torrente, Moreno-Ger, & Fernández-Manjón, 2014). Visualisation of data via OpenID Connect+Open Badges.	No evaluation or real-life use of the LA dashboard with students.
Data mining				
Bader-Natal and Lotze (2011)	Grockit analytics system to answer questions about learning +engagement. Five steps framework: Collection, Selection, Analysis, Visualisation and Distribution. Answers questions relevant to all three target users.	The data used will depend on the question asked; the system seems to cater for a wide range of data about the player, the game, the teaching material, the player's performances...	The system is human processed; a question is proposed; queries SQL are designed and views are created. The system allows for hypothesis testing, integrating a tool for randomised controlled experiments.	Experiment data is collected from students in high school and post-college education.
Martin et al. (2013)	Presents LA applied to an online fraction game. The aim is to understand how fractions are learnt. It is mainly aimed at teachers.	The system tracks every mathematically relevant action made by the student.	Visualisation (state diagrams) to show students' learning process. DM to classify the pathways and identify different types of trajectories	24 ten and eleven-year olds played the game for seven weeks.

Gibson and Clarke-Midura (2015)	Used gameplay log files to answer research questions about the relations between score & time spent, predict student performance etc.	Gameplay logs containing all actions performed, scores, duration of gameplay. Authors also calculated the number of time each student played.	Visualisation in tables and graphs and data mining algorithm (machine learning, clustering etc.)	The data used is logged from real gameplays with 1985 students.
(Blikstein, 2011, 2013)	Multimodal learning analytics are presented and used in order to classify and cluster students learning processes. Mainly aimed at researchers.	Multimodal data collected, snapshots of the code produced, entire portions of text collected for text mining, objects and body movements were tracked using cameras and sensors.	DM was mostly used: Classification according to the learning profiles, Clustering of student progress, Expectation maximisation for text mining	A study was performed for each mode described in the paper.
(Greller, Ebner, & Schön, 2014; Schön, Ebner, & Kothmeier, 2012)	Applies LA in the context of learning multiplication tables. LA is used to adapt the application and give insights on the learning process and students' learning styles.	Three types of data collected: difficulty of the question asked, answers given by the student and total number of questions answered Competence level calculated based on this.	Learning rate displayed as a line graph and DM done manually to discover patterns in learning curves and identify students at risk. Heat maps represent the most difficult questions.	A research study was carried out with 42 primary school pupils in 2011 and another one involving 6000 pupils and over 100 teachers in 2013.
Piech, Sahami, Koller, Cooper, and Blikstein (2012)	Addresses *how* students learn to program. The system models the learning progress and uses this model to predict performance. Mainly aimed at the computing teacher	Snapshots of the code with a timestamp are stored every time a student's work is compiled or saved. Midterm, assignment scores and time spent also collected.	Student's progress is modelled as a Hidden Markov Model (HMM). A K-mean algorithm is then applied to cluster the paths students took through the HMM.	The data from a computing class. They created a model during a semester and verified it with students who took the course over the summer.
Other				
Serrano-Laguna et al. (2016)	The authors propose a model to standardise the collection of data for LA in GBL. The model is based on the tracking of events using a json format (Experience API).	The authors divide a user interaction in 5 parts: 1) a timestamp, 2) a user id, 3) the action (i.e. type of interaction performed), 4) a target (i.e. the game element) and 5) an optional value.	not specified	The model was illustrated in a case study with a simple Q&A Geography game (Countrix).

2.1 Different Types of Data Collected

The first obvious challenge to integrating learning analytics in educational games is deciding what data to collect. The literature review identified five types of useful data game developers and educationalists should consider monitoring when using GBL: time-related data, counts, game actions, scores and player data. These are described below.

- *Time-related data:* Some of the studies identified in the literature monitored data related to time. This can range from the total time spent on an activity (Piech et al., 2012) to the time the player took to perform a particular action or achieve a level, or the time of day the player played (Duval, 2011).

- *Counts:* Some of the systems monitored data in terms of numbers. In his paper, Duval (2011) collected the number of logins and assignments finished while Holman et al. (2013) also collected the number of content views and Greller et al. (2014) the number of questions answered.

- *Game interactions / actions:* This type of data gives an insight into the player's actual interactions with the game. It can be very general, such as a player's state in a game (Serrano-Laguna & Fernandez-Manjon, 2014) or more specific such as clicks or answers given to a question (Greller et al., 2014; Martin et al., 2013; Serrano-Laguna & Fernandez-Manjon, 2014). Piech et al. (2012) even describe how they logged snapshots of code whenever a program was saved or compiled.

- *Scores:* The scores of a player are a very important and relevant measure. The performance of the player can be monitored (Bader-Natal & Lotze, 2011; Reese, 2014; Serrano-Laguna & Fernandez-Manjon, 2014) as well as the badges he/she earned (Holman et al., 2013). Score can also be associated with time to visualise its evolution throughout the gameplay and across gameplays.

- *Player data:* In order to refine the data collected, it is useful to have information about the user. The information can be demographic (e.g. age, gender, language), academic (Serrano-Laguna & Fernandez-Manjon, 2014) or technical with system configuration and session data being logged (Fulantelli et al., 2013).

2.2 Types of Data Analysis

Once the data is collected, an analysis process is needed in order to transform it into useful information. There are two different techniques that could be used: Information Visualisation (IV) that describes the data and Data Mining (DM) that makes predictions based on more complex algorithms.

2.2.1 Information Visualisation (IV)

According to Card, Mackinlay, and Shneiderman (1999, p. 7), IV is "*the use of computer-supported, interactive, visual representations of abstract data to amplify cognition*". Card (2003, p. 211) defines its aim with the following analogy: "*The purpose of information visualization is to amplify cognitive performance, not just create interesting pictures. Information visualizations should do for the mind what automobiles do for the feet*". IV is a tool for humans to draw conclusions about the data available and the visualisation process could be described in six key steps: (i) Mapping – how is information visually encoded? (ii) Selection – among the data available, what is relevant to the considered task? (iii) Presentation – how is the visualization laid out on the available screen space? (iv) Interactivity – what tools are provided to explore and rearrange the visualization? (v) Human factors – are human perceptions and cognitive capabilities being taken into account? (vi) Evaluation – has the effectiveness of the visualization been tested on users? (Chittaro, 2006).

IV can, thus, be seen as a very useful tool to display complex information, such as game data, to a variety of audiences, including players and teachers. It is, by essence, limited and humans are required to draw conclusions from the data and information visualised.

2.2.2 Data Mining (DM)

DM is a tool that draws conclusions automatically from the data collected without requiring the human mind. Rather than being two very different notions, IV and DM can be seen as different levels of data

analysis, IV being a subset of DM but limited to statistics and visualisation and DM providing additional functionality. Data mining is defined by Han, Pei, and Kamber (2011, p. 8) as "*the process of discovering interesting patterns and knowledge from large amounts of data*", they outline that DM is interdisciplinary, using techniques from various other fields such as machine learning, statistics, database systems etc. There are six main tasks performed by DM (Fayyad, Piatetsky-Shapiro, & Smyth, 1996): (i) *Classification* – assigns a class to an object based from its attributes; (ii) *Regression* – real life prediction for an object based on its attributes; (iii) *Clustering* – finds groups (clusters) to categorise the objects; (iv) *Summarisation* – more descriptive, it provides a summary of the data; (v) *Association* – Reveals dependencies between the objects; (vi) *Anomaly detection* – find changes based on previously collected data. A summary of the differences between IV and DM is presented in Table 2.

Table 2: Differences between IV and DM

Information Visualisation	Data Mining
Describes the data collected	Makes predictions based on the data collected
Based on statistics and visualisation	Based on statistics, machine learning, neural networks etc.
Humans are required to draw conclusions	Conclusions are drawn automatically
Low computational cost	Higher computational cost
Accessible to a variety of different audiences	Mostly aimed at experts

2.3 Three Different Target Users

The literature review shows three main perspectives for an LA tool and three different target users associated with them. First, the player perspective allows a player to answer the question "*How am I doing?*". Eight of the papers identified in the review presented a tool aimed at the learners, mainly using visualisation techniques to represent the performance of the player and compare it to the other students in the class or other players of the game. Second, there is the teacher perspective. Educators are the most popular target user in the literature with 14 of the relevant studies offering a solution for them. This perspective allows the teachers to answer the question "*How are my students doing?*". The studies identified use both information visualisation and data mining to present the performance of a group of students, identify students at risk and infer new data such as a prediction of the final grade of the students (Piech et al., 2012). Finally, the least common perspective with three of the papers mentioning it is the researcher's perspective. This perspective allows for further analysis to be made on the games and gameplay. It answers the question "*How are the games used and how are they useful?*". Data mining algorithms are usually used in this perspective.

3. Proposed Approach: EngAGe

Section 1 identified two key problems related to educators' use of GBL as an assessment tool. First, the assessment is embedded into the game's code and educators cannot modify it to suit their students' needs. Second, there is a lack of detailed reports about the students' interactions during the gameplays and their assessment. Authoring tools (Hainey & Connolly, 2013; Torrente, Del Blanco, Marchiori, Moreno-Ger, & Fernández-Manjón, 2010) partially addressed these problems by bypassing game developers and allowing educators to create and modify educational games themselves. This section presents our approach to addressing these problems in a situation where a game is created by a development team. The solution is based on externalising the assessment from a game's mechanics. The modularity allows educators to modify and adapt it to their students' needs, even after the end of the development process and without having to

edit the source code. This functionality addresses the first research problem. It is achieved through an online visual editor. Educators can also visualise the list of their available games along with their unique versions and manage their students' access to them. Having this control over the games will help educators develop a sense of ownership and trust towards the tool. To address the second problem, EngAGe includes an LA dashboard; all the data processed by the engine will be stored allowing educators and researchers to visualise data about all the gameplays of a particular game and across games. The aim is to make LA more accessible to teachers and help them make informed decisions about the changes needed to adapt their games to improve student learning.

EngAGe's interface was designed and developed with the aim of providing teachers with a tool to manage their games, their students, and to visualise gameplay data. The interface's architecture is represented in Figure 1 and includes:

- *A management system for students and games*: From the web interface, educators can create student (player) profiles, group them in classes and give them access to specific versions of the games.
- *An assessment editor:* Based on the information provided by the LA dashboard, teachers are able to make informed decisions about the modifications needed to the games. These modifications are performed through the editor.
- *A learning analytics dashboard*: This dashboard allows educators to visualise and interact with the data collected during the various gameplays. Data mining was included in the design to detect patterns and anomalies.

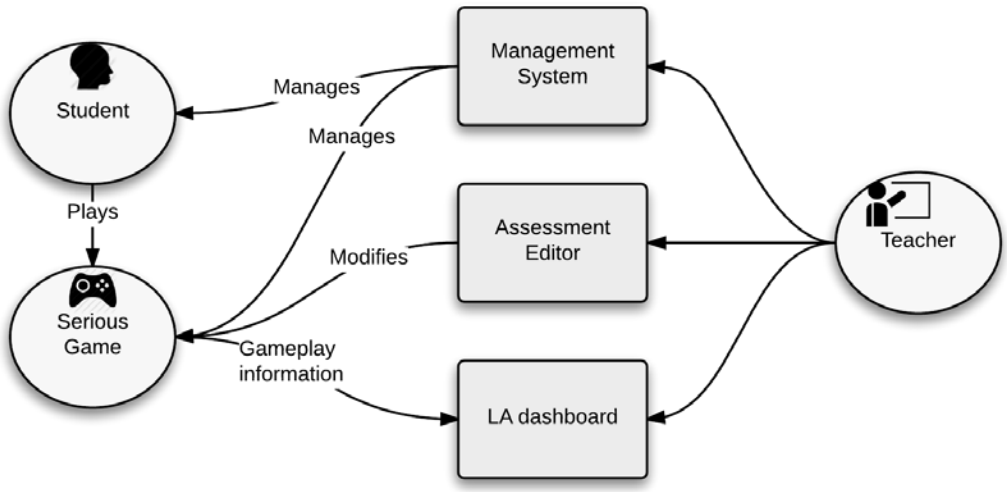


Figure 1: EngAGe web interface

3.1 Games supported by EngAGe

EngAGe supports a variety of serious games. Previous literature reviews (Chaudy & Connolly, in press) suggest that there are five types of assessment possible in serious games: Quizzes, Quests, Monitoring of states, Use of probabilistic model, and Peer assessment. At the time of writing, EngAGe can support the first four types. Peer assessment is not yet fully supported as the system is only able to update one player's score and send him/her feedback: for EngAGe's integration to be successful the game must be synchronous and update a player's score and feedback from the side of the player receiving the assessment. In terms of games genre, most are supported with the exception of asynchronous multiplayer, EngAGe however is

limited to games that have an internet connection (e.g. online games or mobile games on devices connected to the internet) in order to perform the web services calls. More technical details about how EngAGe is integrated into serious games is discussed in another paper (Chaudy & Connolly, submitted).

3.1.1 A mini game to demonstrate the potential of EngAGe

To use as a proof of concept, a mini game was developed using EngAGe. *EU mouse* is an endless runner type game where the player is a mouse running through a geography classroom. The mouse must collect the countries that form the European Union (EU) scattered in the room. The player is given three lives and loses one when collecting a country that is not part of the EU. The game keeps track of the countries found and the player wins when he/she collects all 28 correct ones. The games include three types of feedback. First, invasive messages are shown in a feedback panel stating whether the country selected is indeed part of the EU. Then, final feedback is send when winning or losing the game, this triggers the end of the gameplay. Finally, non-invasive adaptation feedback can be triggered based on the player performance, the speed of the game is automatically adjusted, slowing down when a player has only a life left and speeding if they are performing well. Figure 2 shows a screenshot of the gameplay. The game was distributed online via email, forums and social media. The engine recorded 378 gameplays and 33 players. The data collected will be used to illustrate the assessment editor and the LA dashboard detailed below.

Figure 2: EU mouse gameplay

3.2 An assessment editor

When a game is created using EngAGe, educators have access to all EngAGe tools. This section will present the editing tool that was developed for educators to modify a game.

3.2.1 Design Choices

Based on a literature review performed in 2014 (Chaudy, Connolly, & Hainey, 2014), there are three main designs for educator's GBL authoring tools:

- Text-based user interface: The user is asked to describe the game using text only. This can be done using a Domain-Specific Language (DSL) or existing configuration languages such as eXtensible Markup Language (XML).
- Form-based user interface: The user edits or creates the game using forms. Text fields, drop-down lists, upload buttons etc. are used.
- Visual Editor: The user creates a visual representation of the game. This can be achieved using drag-and-drop options with flow or state diagrams.

The first option was only found in one paper (Rodriguez-Cerezo, Gómez-Albarrán, & Sierra-Rodríguez, 2013) and the tool discussed was aimed at engineering teachers. A text-based interface was not considered for EngAGe as its target users are not restricted to teachers familiar with programming languages. In order to be as inclusive as possible, more user-friendly options were considered. The initial design published in the previously cited paper, included forms for basic information and a visual representation of the assessment logic. However, an expert evaluation carried out with a prototype of the editor concluded that a visual editor might be too complex and difficult to use. The interview resulted in a new design based solely on forms and shaped around a "rules of the game" template that would be more understandable for educators. An online survey with 27 educators was used to confirm the new design. Out of the 27 answers received in the poll, 15 (55.6%) chose the second design and 12 (44.4%) chose the initial design. The difference between the two groups is not statistically significant ($p > 0.05$) and therefore could not be used to decide which design to use. It was decided to use the expert's input and the form-based editor was elected to be fully developed. This choice also allowed for more flexibility as it is easier to make modifications to a form than to a visual language.

3.2.2 Modification Options

Teachers can modify any game they have access to. All the modification features of the editor are detailed below and illustrated with the *EU mouse* game. The assessment logic of a game is based on three main components: scores, feedback and actions. An action can trigger feedback and update scores and scores can trigger feedback when they reach a threshold value. As these components are so closely related, there is some redundancy in the editor.

Game and player

An educator can modify the name and description of the game and can also change the target age range of the game and its keywords. If the game was initially made public by the developer, educators can specify whether they want their newly created version to be public, allowing other teachers using EngAGe to see and use it. For LA purposes, a teacher might want access to specific data about the players such as their age, level or mother tongue. In this first section of the editor, they can update the players' characteristics, adding new ones and updating or deleting existing ones. The new version of the educational game will ask the students to enter the information the system does not already have.

Scores and learning outcomes

The definition of learning outcomes is crucial in a teaching and learning process. A teacher might need to split a general learning outcome into more specific ones in order to identify more precisely where his/her students are struggling. The learning outcomes of the game, and the other scores (e.g. lives, money, ammunitions) can be fully modified in the editor. Feedback associated with the score is also shown and modified in this section. Each score is displayed in a minimalistic way, showing only its name and description but each pane can be extended to show more details as presented in Figure 3. The visual editor allows the creation, modification and suppression of any of the game's scores.

In the EU mouse example, a teacher could disagree with the initial choice to have the *eu_countries* score starting at 28, choosing instead to have it start at zero and not divulge how many countries there are in total. If the gameplay data shows that their students lose the game too often, educators can decide to give players more than 3 lives to start with. The adaptation feedback slowing down or speeding up the game can be adapted to the players' performance; the thresholds for triggering this feedback can be lowered or increased. New feedback messages can also be created to send more information to the player.

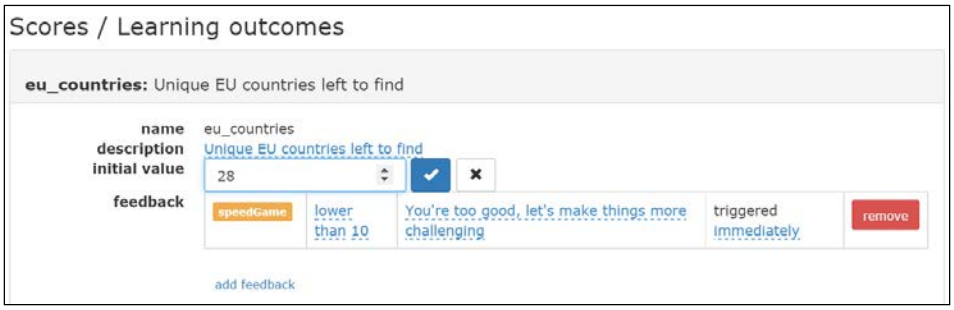


Figure 3: Modifying the learning outcomes and other scores

Evidence model: rules of the game

The evidence model presents the logic of the assessment and how observations provide evidence about the learning or lack thereof. This section defines the meaningful actions of the game from an assessment point of view, describes how the scores are updated after them and what feedback is triggered as a result. The actions of the evidence model are displayed in accordion panes that can be extended or collapsed. New actions cannot be created as they are too closely linked with game mechanics; existing actions are associated with specific functions in the source code and there would not be a method implemented to handle a new action. However, teachers can update existing actions, they can modify the values accepted by the action, how they update the scores and the feedback that they trigger. Many countries want to enter (or leave) the EU. If the list of correct EU countries changes, the EU mouse game would become obsolete. Educators can modify that list in the *rules of the game* section of the editor and make sure that the game's assessment is up to date and that the game can continue to be used. In this section, educators can also create more detailed feedback. For instance, the simple correct and incorrect existing feedback could be extended, and a confirmation message could be given along with the date of entry of the country selected for more elaborate feedback. Figure 4 presents one action of the editor's evidence model.

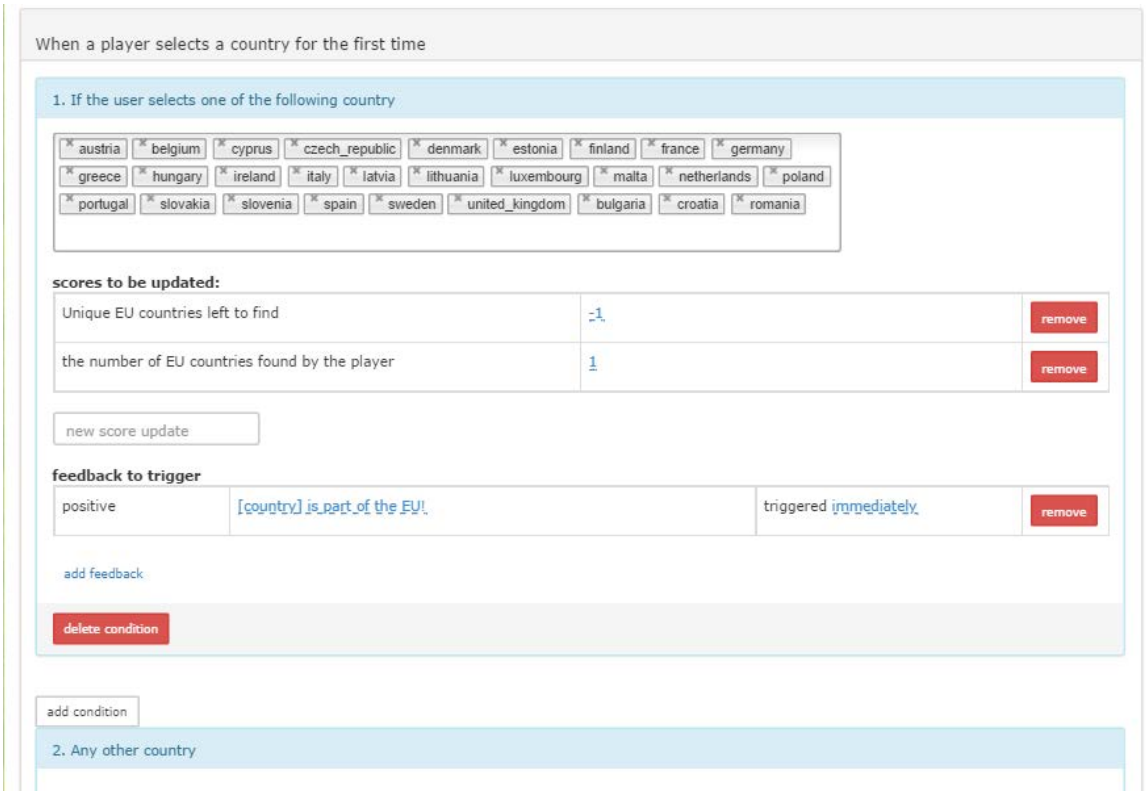


Figure 4: Modifying the evidence model

End of the game

A specific section was created in the editor for the feedback triggering the end of the game. In some cases, the end is triggered by the game mechanics, but if it is linked to the assessment then the educators can modify it fully. They can change the condition and the feedback message. Games can be won, lost or ended without win/lose states. In the EU mouse game, the game is won when all 28 EU countries are found, and it is lost when the player runs out of lives. An educator might want to change this logic and create a practice game without lives, when the player would neither lose nor win and would have to find all 28 countries through trial and error, in which case the game ends without a winning or losing state.

Badges: across gameplays feedback

At the end of the form, educators can visualise the badges. Currently the descriptions of existing badges can be modified as well as the conditions required for earning them, however, new badges cannot be added as they are associated with an image in the game.

3.3 A learning analytics dashboard

Section 1 noted that educators and researchers lack reports about gameplays. Monitoring students' learning through computer games is an arduous task. With many games, the only option to know what is happening during a gameplay is through observation and this is not often possible. An LA dashboard was integrated into EngAGe to allow educators to visualise gameplay data and to monitor their students' progress at any point after the game's distribution. After reviewing the learning analytics, educators can make informed decisions about any changes required to a game's assessment.

3.3.1 Model for the LA dashboard

The model for this LA dashboard is based on the two frameworks found in the literature review (Bader-Natal & Lotze, 2011; Serrano-Laguna et al., 2012) and is presented in Figure 5. This model corresponds to the teacher perspective. It takes into account every gameplay of every game and by every student who is associated with the teacher. It looks across games but is limited to the teacher's students. It is represented as a cycle rather than a pipeline to reflect progress and improvement. The model is composed of six phases:

- *Selection:* Teachers select the data they believe relevant for collection. It can be of any of the types outlined in the literature review (i.e. time, counts, game interactions, scores and player data).
- *Collection:* The teacher's students play the game(s) and all the data previously selected is logged.
- *Analysis*: The data collected in the gameplays is aggregated and analysed using DM techniques. It is then formatted for output to the educator.
- *Visualisation*: Using IV techniques, information such as time spent on games and average scores are displayed along with anomalies detected and predictions.
- *Interaction*: The teacher can interact with the information; refine it by, for example, gameplay, learning outcome, student or class.
- *Reaction*: Based on the information shown, the teacher can now act changing the data collected or adapting the game's assessment.

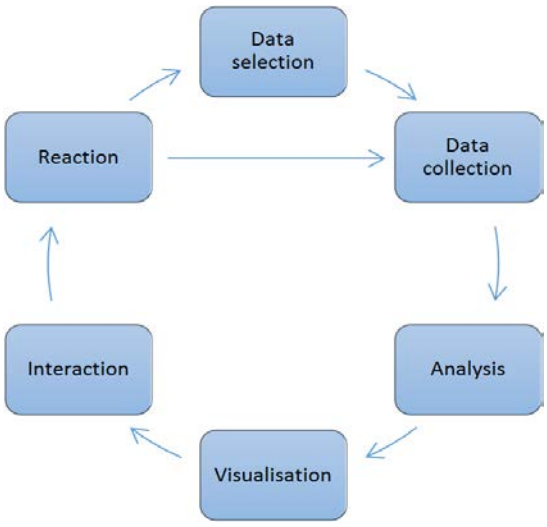


Figure 5: Model for LA in educational games - the teacher's perspective

3.3.2 Dashboard Charts

The LA dashboard design comprises a number of different blocks grouped into four categories: basic information, scores, interaction and data mining. The data mining blocks would require larger quantities of data to be perfected; they will be implemented in a future version of this work. The features available and their organisation were derived from an online survey carried out with 27 teachers. Each block is displayed in a specific colour to allow for easy identification of relevant information and can independently be downloaded, allowing teachers to use the graphs for reports or feedback to stakeholders. The same game can be used by different teachers and when loading an LA dashboard, a teacher will only see the gameplay data related to their own students. Developers and researchers with an administrator's account are not

restricted and can visualise the entire dataset although it is anonymised; the links between players and students are not shown.

EngAGe includes an option for the games to ask a player for some information at the start such as age or gender. All the graphs displayed in this dashboard can be filtered according to this information.

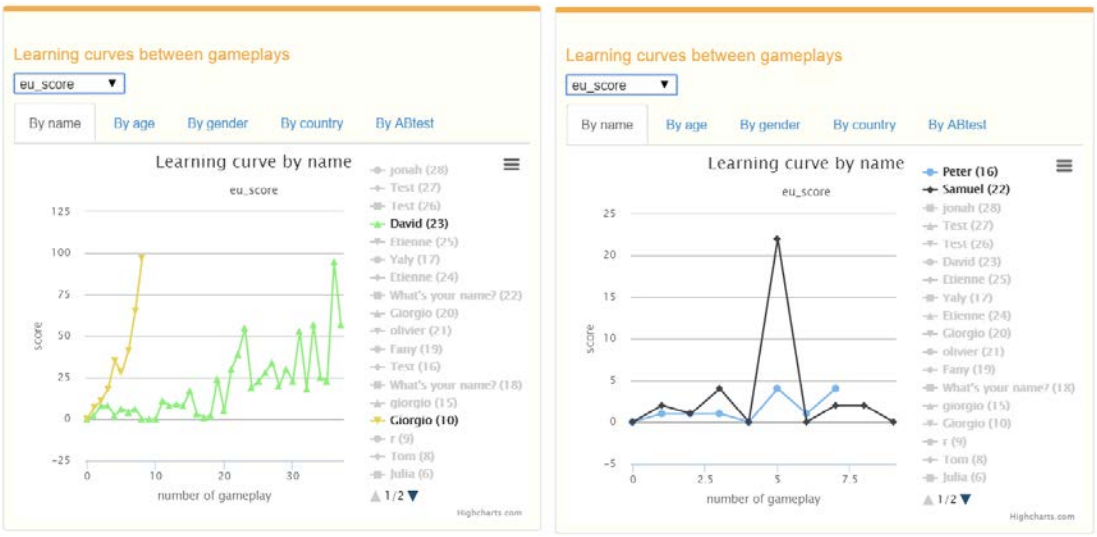
Basic information

The LA dashboard gives basic information about time and numbers. First, a summary of the game's use allows educators to view the number of gameplays, the number of players and to quickly determine whether all their students played. A player can be excluded from the analytics using checkboxes, which is particularly important for teachers testing the game as they will not want their own gameplay data to appear in their class analytics. Then, five blocks offer visual representation of basic information. A pie chart shows the distribution of gameplays by player characteristic and gives an overview of who played the game. A bar chart shows the number of times each student played the game, and another displays the total time spent playing the game, in minutes. These charts are useful to identify students, or groups of students, who played significantly less or more than their peers. The last charts allow teachers to see the day and the time of the day the gameplays took place.

Score visualisation

The LA dashboard also gives information about the player's scores over time. Three blocks were created for this purpose. First, a box plot displays the minimum, maximum, median and quartile values for each score obtained in the game. This information can be filtered by player characteristic (e.g. age or gender) and the box plot can either be based on the entire gameplay data or limited to the first, best or last gameplay for each player. This chart is useful to identify students, or groups of students, in difficulty or to compare different groups.

Next, a line chart provides a visual representation of the learning curves of the students. The chart shows, for each score, the evolution of its value over the different gameplays. This block allows teachers to identify players learning rapidly, steadily or even not learning. Anomalies such as random playing could also be detected. To a certain extent, cheating could also be highlighted by this chart: if a player asked someone else to play in his/her place or used a set of answers once, his/her maximum score would be high and he/she would not have been identified in the previous block, but the learning curve would show an abnormal peak that could be easily spotted. Figure 6 shows the learning curve for different students.



(a) **(b)**

Figure 6: Learning curve of different students: a) Giorgio learning rapidly and David learning steadily, b) Peter not learning and Samuel who might have cheated once

In some cases, the learning curve for a single gameplay can be useful; for example, if the game is to be played only once or for further information about idle time and actions. A third block shows, for each score, the evolution of its value within a gameplay. The chart can help visualise the student's learning style and identify students needing help such as the ones reaching a plateau or demonstrating long idle times. Figure 7 shows the block for the EU mouse game and two students Giorgio and David who only found 26 and 22 EU countries, respectively.

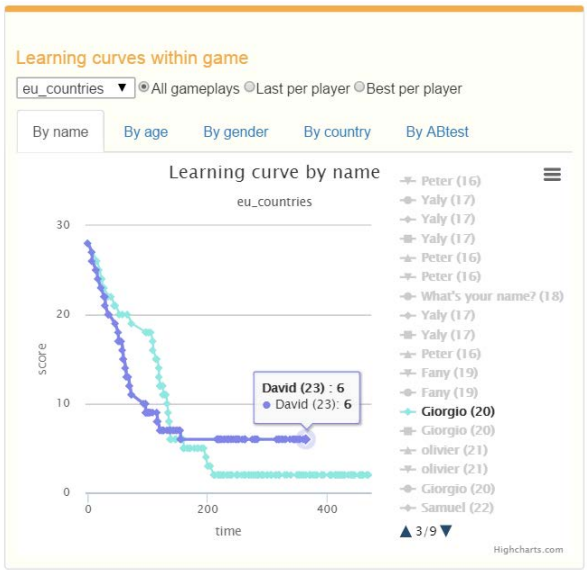


Figure 7: Evolution of the '*eu_countries*' score within a gameplay

Educators often need to manipulate the data themselves. As creators of the assessment, they know the meaning of each score and their relationship to one another. In some cases, scores need to be added to provide meaningful information, or a more complex formula can be required. For example, in a translation game monitoring the words translated correctly and incorrectly, each score taken separately will not be

sufficient for a teacher to verify that the student achieved the learning goals. However, dividing the number of correct answers by the sum of correct and incorrect answers will provide the teacher with a percentage of correct translations carrying more meaning for the teacher. The engine cannot be expected to understand the semantics of the scores and therefore cannot generate the formula itself. Therefore, the previously discussed blocks of the LA dashboard include an option for educators to define a formula using the game scores.

Interaction visualisation

The green blocks of the LA dashboard give more detailed information about the interactions between the players and the game. First, a bar chart shows the actions most or least performed by players during various gameplays. The educator can visualise the evolution of actions between the first, last and best gameplays of each player. For example, Figure 8a shows that 13.79% of the players made the error of selecting 'Serbia' during their first gameplay, but that number is down to 3.45% for the last gameplays, illustrating that they learnt that this country is not part of the EU. Figure 8b also shows the correct EU countries least commonly selected and how they compare between gameplays. A second block offers the possibility of focusing on a specific parameter, for example, a country that did not appear in the previous charts.

(a) **(b)**

Figure 8: Interaction graphs showing a) the most common errors of the EU mouse players, b) the EU countries least found

Similarly, the LA dashboard provides a feedback block displaying, for each piece of feedback, the average number of times it was triggered in a gameplay. The evolution can also be monitored between the first, best and last gameplays of each player and it can help confirm that the students are learning. Figure 9 illustrates the feedback chart with the feedback of the EU mouse game. An improvement can be seen in the adaptation feedback; the number of students displaying sufficient knowledge and triggering an acceleration of the game has significantly increased between the first gameplays and the best ones. This chart also highlights that winning the EU mouse game is extremely difficult, with only 2% of all gameplays and 7% of the best gameplays being won.

(a) **(b)**

Figure 9: Feedback block showing the average number of times a) adaptation feedback and b) final feedback was triggered in the EU mouse gameplays

The last blocks focus on the badges earned by the players. With a radar chart, the teacher can visualise which badges are commonly or uncommonly earned. This chart allows the educator to identify badges that are too easy or too hard to earn. Another block allows teachers to select a badge and access the list of students who earned it and the ones who did not.

Data mining

The LA dashboard design includes two data mining blocks. In contrast to the previous blocks that provide a simple visualization of the data in the database, the data mining ones are based on algorithms that are capable of inferring new data and making predictions. Figure 10 shows the output of these two algorithms for the EU mouse game. A first block uses a K-means clustering algorithm to propose a categorization of the teacher's students based on time spent on the game and the number of gameplays it took students to achieve a target score set by the educator. This information allows teachers to identify different learning styles, which could prove to be useful if they want to divide the class into groups, for tutorials for examples. The clusters can also be used to assign different versions of a game to students. In the example of the EU mouse game, three clusters are identified: cluster 3 correspond to the students who never achieved the educator's goal, cluster 2 to those who achieved it in less than 5 gameplays, and cluster 1 to those who achieved it in more than 15 gameplays. The educators could use this data and offer extra materials or support to the students from clusters 1 and 3 who struggled the most. Three new versions of the game could also be created: a version where the player is given more lives for students in cluster 1, a slower version for students in cluster 3 and a more difficult version (e.g. one life only, faster or more difficult winning state) for students in cluster 2.

The second block uses a polynomial regression algorithm to make predictions. It is based on all the data gathered (time spent and score achieved for instance) and determines the norm for a new gameplay. This tool also allows teachers to identify students who are well above or below average based on their distance to the norm. The interpretation is left to the educator. In the EU mouse example, this blocks shows the expected score based on the time of a gameplay. By using data from both blocks educators can better identify students at risk. Teachers have to be warned that the data inferred by both blocks highly depends on the quality of the data gathered and numbers of gameplays logged. As many of the EU mouse game players only played the game once, the data inferred is not very significant or precise, the data mining blocks would need to be tested in a real-life setting with a class of students in

order to be properly evaluated and adapted. The algorithms proposed here are only showing the potential of the tool.

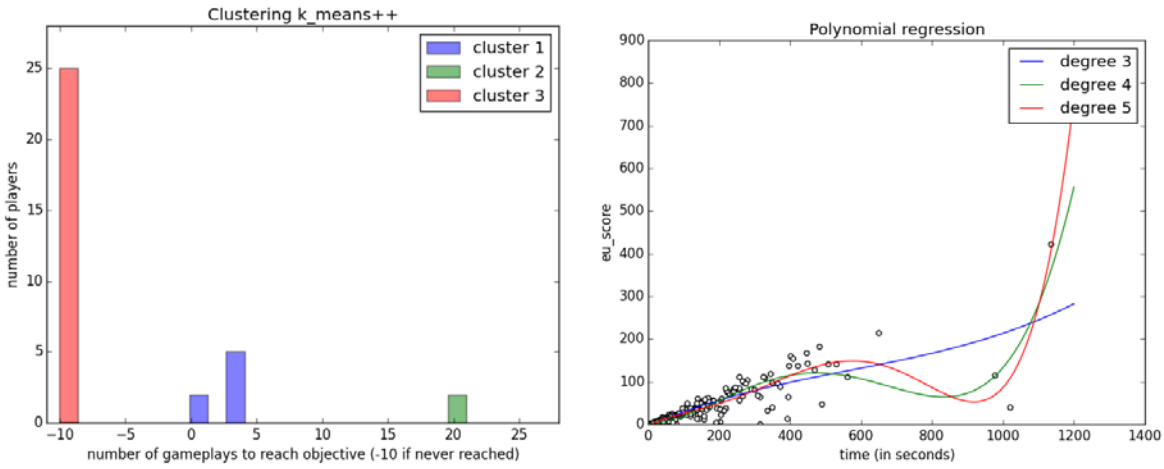


Figure 10: Data mining outputs: clustering and prediction

3.3.3 Players Reports

The dashboard provides information about the groups of students who played the game and is designed to help educators visualise the overall performance of the class. However, teachers often need to access more detailed information about each student. In the '*player reports'* view of the dashboard, educators can create a template of a player's report selecting which charts and numbers are to be included. A player report focuses on the student performance, his scores and badges earned. If the teacher chose to, the report can compare a player's results to others'. Once the template is ready, the teacher can download the report (in PDF form) of a particular player or a class.

4. Evaluation of the tool

This section presents the evaluation of the engine with educators. A user study was carried out to collect educators' opinions on EngAGe for visualising LA through the dashboard and modifying a game's assessment using the editor. A summative evaluation was performed with 31 educators from various backgrounds to evaluate the tool's usability and usefulness. The evaluation also allowed preliminary conclusions to be drawn on the educators' change of opinion regarding GBL as an assessment tool before and after using EngAGe.

4.1 Methodology

This study answers two main research questions (RQ):

1. Can this engine be used effectively by teachers to visualise learning analytics and adapt a game's assessment?
2. Does the engine increase educators' use of GBL and their trust towards a game's assessment?

The evaluation of EngAGe with educators followed a pre-test/post-test pre-experimental model with no control group. In the first questionnaire educators were asked about their opinions of games as a tool for teaching and learning, listing advantages and limitations of GBL and how they would use it. The last question is about GBL as an assessment tool: "*All of your students achieved a good score in a game related*

to your lesson. Would you consider the notion learnt and move on to the next lesson?" Educators had to select one of nine predefined answers between "*No, I would still give the students a paper-based test*" (1) and "*Yes, I would totally trust the game*" (9). After the questionnaire, educators were given access to two educational games created for that purpose; the EU mouse game presented earlier and a language game, vocShoot (Chaudy, 2015) where players protect their planet against meteorites by translating words. Four students were automatically created for them along with several gameplays, this allowed sample data to be shown to the educators. All the participants saw the same data until they used the games with their own students or played themselves. A tutorial took educators through the process of using EngAGe to create new students, visualise LA and modify a game's assessment. They were then given the opportunity to use the games and tools freely; they could customise the games to suit their class, use the game with their students, use the LA dashboard…

At the end of the study, educators answered the post-questionnaire comprising of (i) the same two questions about the use of GBL in the classroom and as an assessment tool; (ii) questions about the usability and usefulness of EngAGe interface; (iii) questions about the usability and usefulness of the LA dashboard; and (iv) questions about the usability and usefulness of the assessment editor. The educators' opinions on usability were measured using the System Usability Scale (SUS) (Brooke, 1996) that has been proven to be robust and versatile (Bangor, Kortum, & Miller, 2008).

The data gathered in this user study will be analysed using both descriptive statistics and non-parametric tests. A sample size of 31 was calculated using Xu's formula for descriptive studies in quantitative research (Xu, 1999) and Eng's formula for non-parametric tests (Eng, 2003).

4.1.1 Participants

This study was distributed online via email, precautions were taken to avoid unreliable answers related to online distribution of the questionnaires. Educators were asked to provide an email address in the first questionnaire, use it to create an EngAGe account and specify their EngAGe username in the post questionnaire. This way the authors could: (i) identify the related pre- and post-questionnaires; (ii) ensure that a teacher only participated once; and (iii) verify that the participants actually completed the tutorials (using EngAGe's monitoring tool to visualise game versions created by educators). 31 answers were recorded. Participants included a wide variety of educators. Females were slightly more represented (61%) and 52% of participants already used games for teaching. The teaching experience of participants ranged from 0 to 30 years with a mean of 9.5 years and a standard deviation of 8.1. The participants' age, nationality, subject taught and level taught are summarised in Table 3.

Table 3: Distribution of the participants

	Category	Participants	Percentage
Age	between 19 and 25	5	16%
	between 26 and 35	15	48%
	between 36 and 50	9	29%
	between 51 and 65	2	7%
Nationality	United Kingdom	12	39%
	France	5	16%
	Bulgaria	3	10%
	Romania	2	7%
	Germany, Greece, Indonesia, Ireland, Malaysia, Nigeria, Pakistan, Poland and Spain	1 each	3% each
Subject taught	technology	8	26%
	science	6	20%
	social studies, mathematics	5 each	16% each
	foreign language	4	13%
	Literature, Arts, Business	1 each	3% each
Level taught	higher education, further education	15	47%
	Secondary, primary, pre-primary	16	53%

4.2 Games and Students Management System

The first section of the tutorial focused on reviewing the available games and creating new students. Educators were also asked to use the student-game access table to associate students with versions of the games to play. The post-questionnaire asked teachers to reflect on usability and usefulness of the student and game management system rating its usability and usefulness. The usability was rated by educators using five-point Likert scales between '*Very difficult to use'* and '*Very easy to use'*. The findings show that most participants (94%) found the interface easy to use, 72% rating it as '*very easy to use*'. The usefulness was rated by educators using a five-point Likert scale between '*Least useful'* and '*Most useful'*. The results suggest that all participants found the interface useful, with 74% rating it most useful. These results are confirmed by the fact that all participants attempting the tutorial completed the first section (i.e. creation of students and selection of game versions to play). These findings suggest that EngAGe's interface can be used effectively by educators for visualising the games they have access to, modifying the list of their students and for assigning students to certain versions of the games to play.

4.3 The LA dashboard

To answer the first part of RQ1, the LA dashboard was evaluated using two grid questions and an open question for comments. The usefulness of each of the LA charts and the player reports were evaluated using a five-point Likert scale ranging from '*Absolutely not useful'* to '*Extremely useful'*. Figure 10 presents the answers on diverging stacked bar charts as recommended by Robbins and Heiberger (2011). The answers were then graded on a corresponding scale from 1 to 5. Table 4 shows the mean ratings for each element of the LA dashboard ranking from highest to lowest. The findings suggest that all the charts in the LA dashboard are useful, with the best rated elements being the players' reports (4.81, SD = 0.39) and the chart showing the total time students spent playing (4.76, SD = 0.43). The lowest rated chart is the one showing the time of the day students played (4.17, SD = 0.99).

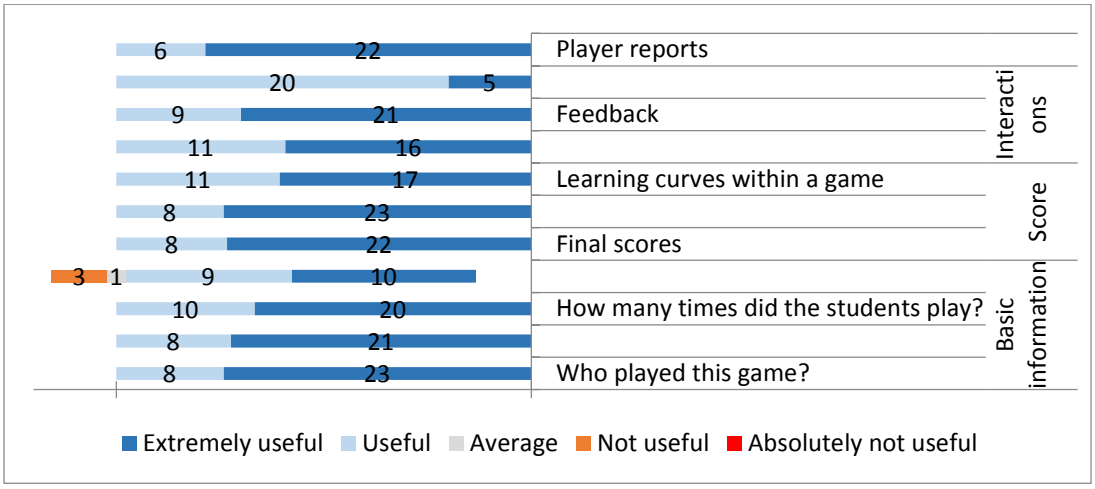


Figure 11: Usefulness of LA dashboard

Table 4: Ranking of EngAGe's LA elements by usefulness

LA element	Mean	SD
Player reports	4.81	0.39
How long was the game played?	4.76	0.43
Learning curves between gameplays	4.74	0.44
Who played this game?	4.71	0.45
Final scores	4.7	0.46
Feedback	4.69	0.46
How many times did the students play?	4.66	0.48
Learning curves within a game	4.65	0.48
Most and least common actions	4.63	0.48
Badges	4.23	0.42
When did the students play?	4.17	0.99

The usability of the LA dashboard was also evaluated using the SUS questionnaire. The mean SUS score for the LA dashboard is 76.3 with a median of 80 and a standard deviation of 19.1. Bangor et al. (2008) propose a seven-point adjective rating scale for representing SUS scores ranging from "*Worst Imaginable*" to "*Best Imaginable*". The LA dashboard score obtained corresponds to an *Excellent* one. Table 5 lists all the SUS statements and, for each of them, the number of participants who elected *Agree* or *Strongly agree*, the equivalent percentage, the mean rating and standard deviation. The results are clearly overall positive; 26 (84%) stated that they would like to use the LA dashboard again and 25 (81%) think that most educators would be able to use the system quickly. Negative statements reflect the possible areas of improvement of the system, such as its complexity and learning curve with 7 (23%) stating that they would need support to use it in the future and 5 (16%) finding the system too complex.

Table 5: Participants' answers to the SUS questionnaire for the LA dashboard

Statement	Participants agreeing	Participants disagreeing	Mean	SD
Positive statements				
I thought the system was easy to use	28 (90%)	2 (6%)	4.1	0.86
I think I would like to use this system in the future when using educational games	26 (84%)	2 (6%)	4.32	0.89
I found the various functions in this system were well integrated	26 (84%)	1 (3%)	4.23	0.79
I would imagine that most teachers would learn to use this system very quickly	25 (81%)	2 (6%)	3.81	0.78
I felt very confident using the system	23 (74%)	5 (16%)	3.87	1.01
Negative statements				
I think that I would need support to be able to use this system	7 (23%)	21 (68%)	2.35	1.12
I found the system unnecessarily complex	5 (16%)	24 (77%)	1.97	1.06
I found the system very cumbersome to use	3 (10%)	25 (81%)	1.81	0.96
I needed to learn a lot of things before I could get going with this system	3 (10%)	23 (74%)	2.06	1.11
I thought there was too much inconsistency in this system	2 (6%)	25 (81%)	1.61	0.94

The questionnaire allowed participants to add qualitative comments. These comments give more information about the complexity of the system and suggest that the dashboard might be displaying too much information at once as many educators found it confusing. Example qualitative comments are:

- *"I found the dashboard quite daunting with so much information being displayed on a single page. May be better to divide this into a number of pages?"*
- *"There was quite a lengthy delay before the figures loaded. Also, page had a lot of information displayed at any one time, which might be off-putting for some teachers."*
- *"The LA dashboard was very busy so some way of displaying less at first and giving the user the option to display more graphs might be helpful."*
- *"Various functions of dashboard could appear after being requested instead of being visible all together at once."*

As a result of this feedback, the LA dashboard was changed and divided into four different tabs: General information, Scores, Action and feedback, and Badges. This modification also allowed for the page to be displayed quicker as each tab was loaded separately.

4.4 The Assessment Editor

The second part of RQ1 focuses on whether educators can effectively adapt a game's assessment using EngAGe (i.e. changing how assessment is performed and how feedback is given in the game itself). EngAGe's editor allows educators to modify seven key assessment elements: game information, player information, scores, feedback, assessment logic (rules), conditions for ending the game and badges logic. Participants evaluated each of these elements separately using a five-point Likert scale ranging from

'*Absolutely not useful*' to '*Extremely useful*'. Figure 11 shows diverging stacked bar charts representing a breakdown of the answers. These answers were then graded on a scale from 1 (*Absolutely not useful*) to 5 (*Extremely useful*) and Table 6 shows the mean ratings ranking from highest to lowest. These very positive findings suggest that educators found EngAGe's assessment editor useful, the most important sections being the feedback and assessment logic (4.79, SD = 0.41 and 4.76, SD = 0.43) and the least important one being the badges logic (4.38, SD = 0.49).

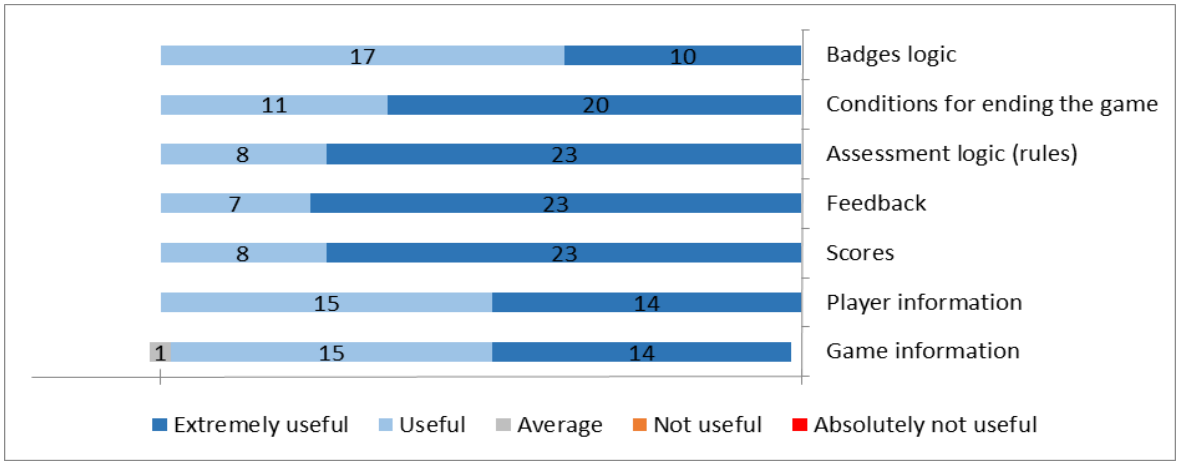


Figure 12: Usefulness of EngAGe's assessment editor

Table 6: Ranking of the functionalities of EngAGe's assessment editor by usefulness

Editor's functionality	**Mean**	**SD**
Feedback	4.79	0.41
Assessment logic (rules)	4.76	0.43
Scores	4.74	0.44
Conditions for ending the game	4.63	0.48
Player information	4.52	0.5
Game information	4.4	0.55
Badges logic	4.38	0.49

The usability of the editor was evaluated using the SUS questionnaire. The mean SUS score computed was 77.7 with a median of 82.5 and a standard deviation of 15.3, which according to the adjective rating scale corresponds to an *Excellent* score. Table 7 lists all the SUS statements and, for each of them, the number of participants who elected *Agree* or *Strongly agree*, the equivalent percentage, the mean rating and standard deviation. The results are clearly overall positive; 29 (94%) stated that they would like to use the system again and none found inconsistencies. However, five participants (16%) stated that they would need support to use it in the future, suggesting that the documentation and overall help to educators could be improved.

Table 7: Participants' answers to the SUS questionnaire for the assessment editor

Statement	Participants agreeing	Participants disagreeing	Mean	SD
Positive statements				
I think I would like to use this system in the future when using educational games	29 (94%)	1 (3%)	4.58	0.71
I found the various functions in this system were well integrated	28 (90%)	1 (3%)	4.13	0.66
I thought the system was easy to use	26 (84%)	2 (6%)	3.97	0.74
I would imagine that most teachers would learn to use this system very quickly	26 (84%)	2 (6%)	3.81	0.74
I felt very confident using the system	25 (81%)	3 (10%)	3.81	0.86
Negative statements				
I think that I would need support to be able to use this system	5 (16%)	22 (71%)	2.32	0.89
I found the system very cumbersome to use	3 (10%)	26 (84%)	1.74	1.05
I found the system unnecessarily complex	2 (6%)	27 (87%)	1.77	0.83
I needed to learn a lot of things before I could get going with this system	2 (6%)	27 (87%)	1.74	0.95
I thought there was too much inconsistency in this system	0 (0%)	27 (87%)	1.65	0.7

Some of the qualitative comments of the participants are included below, the negative ones focus mainly on the complexity of the tool:

- *"It's fantastic that I can change the game without going back to the original developers. This seems a really useful feature to have."*
- *"The editor was quite easy to use; however, I was initially overwhelmed by the amount of information presented. I could see some teachers being deterred by this so some way of simplifying the layout would be helpful."*
- *"Everything was displayed in a single page, which some teachers would be put off at. You might want to divide the page into a number of pages"*
- *"Easy to use once I got started"*
- *"Editing was more complex than other aspects of tutorial and it would be helpful to simplify the interface for less IT-literate teachers"*
- *"All really useful. An excellent tool for teachers if it was available to us"*

4.5 Overall qualitative comments

An optional question asked the participants to list the advantages and limitations of EngAGe as a whole and some of their comments are shown below. Overall, the educators listed more advantages than limitations. Their comments suggest that they understand the potential of the tool and the possibilities it offers for educational games. They stressed how helpful the LA dashboard could be, and how it would help them identifying learning difficulties and patterns. They highlighted the importance of being able to modify

the games themselves. However, they also expressed their concerns about the amount of information displayed on both the LA dashboard and the editor and some possible usability issues for complex games or for educators less comfortable with technology.

Advantages

- *"being able to see how my students are performing makes the use of games in the classroom even more appealing"*
- *"For educators, this tool is very useful to know if the students have problems with one topic, or with one question or aspect of the game. For example, in the European countries game, if most of the players answered that Iceland was part of the EU, then, I know that I should teach my student the right answer. The tool and the charts are very useful. In one hand, to understand the students, and to know their weaknesses. In other hand, to control that they understood and learn their lesson."*
- *"Allows a contemporary approach of learning, based on student daily experience"*
- *"Let's me change games to suit my class. Let's me see how my students are performing and where they are having difficulty"*
- *"I find it amazing how easy it is to administrate and analyse a game. But most impressive is the ability to modify the goals and underlying rules."*
- *"This is a really useful tool and being able to see how my students are performing, it makes the use of games in the classroom even more appealing. Similarly, being able to modify the game is also really useful and means I can modify games that I couldn't have changed previously."*
- *"A tool like EngAGe can allow developers and educators alike to see player trends when they're playing the game and allow for a very adaptive learning environment."*
- *"Ability to edit a game to suit my teaching, Ability to see how my students are performing, Ability to change the game based on how my students are performing"*

Limitations

- *"Too time consuming"*
- *"Perhaps too much detail present on the screen at once. Cognitive overload occurs"*
- *"I found it a bit difficult to amend the game the way I wanted it to be. So additional guidance or tutorial would be needed for teachers so they would comfortably use the tools and modify games according to their expectations and needs for the students."*
- *"I assume that EngAGe is very good for games with a small or medium complexity. However, it might lose its benefit when the games become more complex such as games for business simulation."*

Overall, RQ1 can be answered positively: educators can effectively use EngAGe to visualise learning analytics and adapt a game's assessment. The evaluation suggested that the engine would be both useful and easy to use for them and that they would like to have it integrated to the games they use.

4.6 Change of Opinion towards the Use of GBL

To answer the second research question, two questions were asked in both the pre-questionnaire and the post-questionnaire related to the use of GBL in the classroom and as an assessment tool. A first one asked participants to rate, on a five-point Likert scale, different ways of using GBL: as optional homework, compulsory homework, free time activity, in class to practice a notion and as an assignment. The second question asked them to rate how much they would trust a game's assessment on a nine-point Likert scale from "*No, I would still give the students a paper-based test*" to "*Yes, I would totally trust the game*". The post-questionnaire asked educators to answer the questions again reflecting on their use of EngAGe.

Answers were compared to determine whether having access to EngAGe would change how educators use and trust educational games.

A Wilcoxon matched-pairs signed ranks test was performed and its output is shown in Table 9 with the means for each item detailed in Table 10. Results for using GBL as optional homework, to practice a notion in class, or as a graded assignment do not show any significant increase ($p > 0.05$). However, they suggest that EngAGe increases significantly educators' confidence in using GBL as compulsory homework (increase = 0.33, $Z = -2.153$, $p < 0.05$) or as a free time activity (increase = 0.25, $Z = -2.309$, $p < 0.05$). Finally, results suggest that educators would be more likely to trust a game's assessment if they had access to EngAGe (increase = 1.71, $Z = -3.641$, $p < 0.05$) with a large effect size ($r = 0.46$).

Table 8: Output for Wilcoxon matched-pairs, signed rank test showing change in opinion towards GBL

	Optional homework	Compulsory homework	Free time activity	In class practice	Graded assignment	Trust of GBL for assessment
Z	-1.134	-2.153	-2.309	-.489	-1.812	-3.641
Asymp. Sig. (2-tailed)	.257	.031	.021	.625	.070	.000

Table 9: Descriptive statistics of educators' opinions of GBL

	Pre-questionnaire		Post-questionnaire	
	Mean	**SD**	**Mean**	**SD**
Optional homework	3.65	1.02	3.74	1.06
Compulsory homework	2.77	1.02	3.10	1.30
Free time activity	3.94	.73	4.19	.48
In class practice	3.87	.96	3.94	.85
Graded assignment	2.87	1.09	3.16	1.19
Trust of GBL for assessment	3.87	2.11	5.58	2.66

A Mann-Whitney U test was conducted to compare the increase in trust towards using GBL as an assessment tool for participants based on their gender, age, subject taught, level taught, current use of GBL and experience. The first three comparisons did not show any significant difference. The other three tests highlighted that the increase in trust was significantly higher ($p < 0.05$) for further and higher education (colleges and universities) teachers compared to secondary, primary and pre-primary teachers, for teachers currently not using GBL compared to those already using GBL and for teachers with six or less years of experience compared to those with ten or more. The findings for the three comparisons are listed in Table 11. However, interpretation of these findings is difficult as all three groups showing a bigger increase also start with a higher trust towards GBL for assessment in the pre-questionnaire.

Table 10: Comparison of educators' trust towards GBL for assessment

Variable	Group	n	Trust of GBL for assessment		Increase	Z	p
			pre-questionnaire	post-questionnaire			
Level	Further/Higher education	16	4.8	7.2	2.4	-2.395	0.021

	Secondary/Primary/ Pre-primary	15	2.9	3.9	1		
Use of GBL	Doesn't use GBL	15	4.8	7.3	2.5	-2.520	0.015
	Currently uses GBL	16	3	3.9	0.9		
Experience	6 years or less	17	4.6	6.9	2.3	-2.614	0.012
	10 years or more	14	3	3.9	0.9		

In conclusion, RQ2 can also be answered positively: findings suggest that EngAGe would increase the educator's use of GBL in the classroom in some ways (compulsory homework and free time activity) and that they are more likely to trust a game for assessment if they have access to EngAGe tools.

4.7 Limitations of the study

The user study showed that educators can effectively use the LA dashboard and the editor to visualise gameplay information and adapt a game's assessment to their students. The evaluation was quantitative and involved sufficient participants to guarantee a high level of statistical confidence; the conclusions are therefore relatively reliable. However, the evaluation is only preliminary and there are several limitations to the approach taken. First, the participants were not selected at random; they were all volunteers. Next, because the study targeted a variety of subject domains and levels, it was not possible to create games that were useful for all teachers. The games provided helped educators understand the potential of EngAGe, however, not all of them were able to actually use the games in their classroom.

5. Conclusion and future work

Assessment is a crucial part of any teaching and learning process; it must be carefully integrated in educational games. This paper has presented the research project EngAGe (an Engine for Assessment in Games), its background and motivations. The engine is used by developers when creating educational games resulting in a separation of the assessment from the game's mechanics. As a result of this modularity, educators can modify the game's assessment and adapt it to their players via an online visual editor and retrieve information about the gameplays with an LA dashboard. EngAGe's tools for educators help overcome the "*black box*" issue of educational games by providing detailed reports on the gameplays and the possibility of adapting the game according to the students' needs. The LA dashboard could also be used by researchers to gather empirical evidence on the use of educational games in general.

This study had two main research questions: (1) "*Can this engine be used effectively by teachers to visualise learning analytics and adapt a game's assessment*?" and (2) "*Would the engine increase educators' use of GBL and their trust towards a game's assessment*?". A usability and usefulness evaluation was performed with overall very positive results; both the assessment editor and the LA dashboard were rated useful by participants. The system usability was graded using the SUS: the LA dashboard received a mean score of 76.3 and the editor 77.7 both of which correspond to "Excellent" scores. These findings suggest that EngAGe can be used efficiently and effectively by educators after distribution of a game during the teaching and learning process. An analysis of the difference in participants' opinions toward the use of GBL measured in pre- and post-questionnaires suggests that using EngAGe would significantly increase the trust of educators towards using educational games as assessment tools (1.71-point increase on a 9-point Likert scale, $p < 0.05$).

Future work will include integrating the engine in various existing projects working closely with game developers and educators and collecting their opinion. Data mining blocks showing the output of anomaly detection and prediction algorithms will be added to the LA dashboard and evaluated. A further experiment

with teachers in long-term real-life settings will also be considered to provide confirmation of the findings regarding their increase of trust towards GBL as an assessment tool. Finally, an interesting future direction for this research would be to expand the engine to allow for modification of more aspects of the game not necessarily related to assessment such as content, story line, graphics and sound.

References

Bader-Natal, A., & Lotze, T. (2011). *Evolving a learning analytics platform.* Paper presented at the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction, 24*(6), 574-594.

Blikstein, P. (2011). *Using learning analytics to assess students' behavior in open-ended programming tasks.* Paper presented at the 1st international conference on learning analytics and knowledge, Banff, AB, Canada.

Blikstein, P. (2013). *Multimodal learning analytics.* Paper presented at the 3rd International Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry, 189*, 194.

Card, S. K. (2003). Information visualization. In A. J. Julie & S. Andrew (Eds.), *The human-computer interaction handbook* (pp. 544-582): L. Erlbaum Associates Inc.

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think*: Morgan Kaufmann.

Chaudy, Y. (2015). vocShoot. Retrieved from http://yaellechaudy.com/tutorial_vocshoot/

Chaudy, Y., & Connolly, T. (in press). Integrating Assessment, Feedback and Learning Analytics in educational games: Literature Review and Design of an Assessment Engine. In A. Azevedo & J. Azevedo (Eds.), *Handbook of Research on E-Assessment in Higher Education*.

Chaudy, Y., & Connolly, T. (submitted). *Specification and Evaluation of an Assessment Engine for Developing More Flexible Educational Games*.

Chaudy, Y., Connolly, T., & Hainey, T. (2014). *An Assessment Engine: Educators as Editors of their Serious Games' Assessment.* Paper presented at the ECGBL2014-8th European Conference on Games Based Learning: ECGBL2014.

Chittaro, L. (2006). Visualizing information on mobile devices. *Computer, 39*(3), 40-45.

Cózar-Gutiérrez, R., & Sáez-López, J. M. (2016). Game-based learning and gamification in initial teacher training in the social sciences: an experiment with MinecraftEdu. *International Journal of Educational Technology in Higher Education, 13*(1), 2. doi:10.1186/s41239-016-0003-4

Duval, E. (2011). *Attention please!: learning analytics for visualization and recommendation*. Paper presented at the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.

Eng, J. (2003). Sample Size Estimation: How Many Individuals Should Be Studied? 1. *Radiology, 227*(2), 309-313.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Freire, M., del Blanco, A., & Fernandez-Manjon, B. (2014, 3-5 April 2014). *Serious games as edX MOOC activities.* Paper presented at the Global Engineering Education Conference (EDUCON), 2014 IEEE.

Fulantelli, G., Taibi, D., & Arrigo, M. (2013). *A semantic approach to mobile learning analytics.* Paper presented at the 1st International Conference on Technological Ecosystem for Enhancing Multiculturality.

Gibson, D., & Clarke-Midura, J. (2015). Some psychometric and design implications of game-based learning analytics *E-Learning Systems, Environments and Approaches* (pp. 247-261): Springer.

Greller, W., Ebner, M., & Schön, M. (2014). Learning Analytics: From Theory to Practice–Data Support for Learning and Teaching *Computer Assisted Assessment. Research into E-Assessment* (pp. 79-87): Springer.

Hainey, T., & Connolly, T. (2013). *Development and Evaluation Of a Generic e-CLIL Web 2.0 Games Engine*. Paper presented at the 7th European Conference on Games Based Learning, Porto, Portugal.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Harrer, A. (2013). *Analytics of collaborative planning in Metafora: architecture, data, and analytic methods*. Paper presented at the 3rd International Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Harteveld, C., & Sutherland, S. C. (2015). *The Goal of Scoring: Exploring the Role of Game Performance in Educational Games.* Paper presented at the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.

Herrler, A., Grubert, S., Kajzer, M., Behrens, S., & Klamma, R. (2016). Development of Mobile Serious Game for Self-assessment as Base for a Game-Editor for Teachers. In A. De Gloria & R. Veltkamp (Eds.), *Games and Learning Alliance: 4th International Conference, GALA 2015, Rome, Italy, December 9-11, 2015, Revised Selected Papers* (pp. 71-79). Cham: Springer International Publishing.

Holman, C., Aguilar, S., & Fishman, B. (2013). *GradeCraft: what can we learn from a game-inspired learning management system?* Paper presented at the 3rd International Conference on Learning Analytics and Knowledge.

Johnson, L., Adams, S., Cummins, M., Estrada, V., Freeman, A., & Ludgate, H. (2013). The NMC horizon report: 2013 higher education edition.

Johnson, W. L. (2007). Serious use of a serious game for language learning. *Frontiers in Artificial Intelligence and Applications, 158*, 67.

Ketelhut, D. J., & Schifter, C. C. (2011). Teachers and game-based learning: Improving understanding of how to increase efficacy of adoption. *Computers & Education, 56*(2), 539-546.

Kickmeier-Rust, M. D., & Albert, D. (2013). Learning Analytics to Support the Use of Virtual Worlds in the Classroom. In A. Holzinger & G. Pasi (Eds.), *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (Vol. 7947, pp. 358-365): Springer Berlin Heidelberg.

Kiili, K., & Ketamo, H. (2017). Evaluating cognitive and affective outcomes of a digital game-based math test. *IEEE Transactions on Learning Technologies*.

Liu, M., Lee, J., Kang, J., & Liu, S. (2015). What We Can Learn from the Data: A Multiple-Case Study Examining Behavior Patterns by Students with Different Characteristics in Using a Serious Game. *Technology, Knowledge and Learning*, 1-25.

Lv, Z., Esteve, C., Chirivella, J., & Gagliardo, P. (2017). Serious game based personalized healthcare system for dysphonia rehabilitation. *Pervasive and Mobile Computing, 41*, 504-519. doi:https://doi.org/10.1016/j.pmcj.2017.04.006

Martin, T., Aghababyan, A., Pfaffman, J., Olsen, J., Baker, S., Janisiewicz, P., . . . Smith, C. P. (2013). *Nanogenetic learning analytics: illuminating student learning pathways in an online fraction game*. Paper presented at the 3rd International Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Martinez-Ortiz, I., & Fernandez-Manjon, B. (2017). *Full Lifecycle Architecture for Serious Games: Integrating Game Learning Analytics and a Game Authoring Tool*. Paper presented at the Serious Games: Third Joint International Conference, JCSG 2017, Valencia, Spain, November 23-24, 2017, Proceedings.

Minović, M., & Milovanović, M. (2013). *Real-time learning analytics in educational games.* Paper presented at the 1st International Conference on Technological Ecosystem for Enhancing Multiculturality.

Minović, M., Milovanović, M., Šošević, U., & González, M. Á. C. (2015). Visualisation of student learning model in serious games. *Computers in Human Behavior, 47*, 98-107.

Piech, C., Sahami, M., Koller, D., Cooper, S., & Blikstein, P. (2012). *Modeling how students learn to program*. Paper presented at the 43rd ACM technical symposium on Computer Science Education, Raleigh, North Carolina, USA.

Reese, D. (2014). Digital Knowledge Maps: The Foundation for Learning Analytics Through Instructional Games. In D. Ifenthaler & R. Hanewald (Eds.), *Digital Knowledge Maps in Education* (pp. 299-327): Springer New York.

Robbins, N. B., & Heiberger, R. M. (2011). *Plotting Likert and other rating scales.* Paper presented at the Proceedings of the 2011 Joint Statistical Meeting.

Rodriguez-Cerezo, D., Gómez-Albarrán, M., & Sierra-Rodríguez, J.-L. (2013). *Interactive educational simulations for promoting the comprehension of basic compiler construction concepts.* Paper presented at the Proceedings of the 18th ACM conference on Innovation and technology in computer science education.

Sandford, R., Ulicsak, M., Facer, K., & Rudd, T. (2006). Teaching with games. *COMPUTER EDUCATION-STAFFORD-COMPUTER EDUCATION GROUP-, 112*, 12.

Schön, M., Ebner, M., & Kothmeier, G. (2012). *It's just about learning the multiplication table*. Paper presented at the 2nd International Conference on Learning Analytics and Knowledge.

Serrano-Laguna, Á., & Fernandez-Manjon, B. (2014, 3-5 April 2014). *Applying learning analytics to simplify serious games deployment in the classroom.* Paper presented at the Global Engineering Education Conference (EDUCON), 2014 IEEE.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2016). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces*. doi:http://dx.doi.org/10.1016/j.csi.2016.09.014

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces, 50*, 116-123.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2012). Tracing a Little for Big Improvements: Application of Learning Analytics and Videogames for Student Assessment. *Procedia Computer Science, 15*, 203-209.

Serrano-Laguna, Á., Torrente, J., Moreno-Ger, P., & Fernández-Manjón, B. (2014). Application of Learning Analytics in educational videogames. *Entertainment Computing, 5*(4), 313-322. doi:http://dx.doi.org/10.1016/j.entcom.2014.02.003

Siemens, G., & Gasevic, D. (2012). Guest Editorial-Learning and Knowledge Analytics. *Educational Technology & Society, 15*(3), 1-2.

Sliney, A., & Murphy, D. (2008). *JDoc: A serious game for medical learning.* Paper presented at the Advances in Computer-Human Interaction, 2008 First International Conference on.

Torrente, J., Del Blanco, Á., Marchiori, E. J., Moreno-Ger, P., & Fernández-Manjón, B. (2010). *< e-Adventure>: Introducing educational games in the learning process.* Paper presented at the Education Engineering (EDUCON), 2010 IEEE.

Torrente, J., Serrano-Laguna, Á., del Blanco Aguado, Á., Moreno-Ger, P., & Fernandez-Manjon, B. (2014). Development of a Game Engine for Accessible Web-Based Games *Games and Learning Alliance* (pp. 107-115): Springer.

Xu, G. (1999). Estimating sample size for a descriptive study in quantitative research. *Quirk's Marketing Research Review, 1*.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25-32.